Computer Assisted Cluster Data Analysis to Augment the Student Evaluation for Multiple Choice Type Examinations

Kavita OZA
Shivaji University. Kolhapur, India
skavita.oza@gmail.com

Snehal PATIL
Shivaji University. Kolhapur, India
snehalpatil2311@gmail.com

Rajanish KAMAT
Shivaji University. Kolhapur, India
raj_kamat@yahoo.com

ABSTRACT

Purpose :

As showcased in the present paper, it has the potential to analyze underlying patterns in order to predict the learning outcomes and to formulate the evaluation strategy specifically in case of the multiple choice questions. This in turn is also been used to identify the subject areas where students is weak and prone for remedial teaching.

Design/Methodology :

Computer assisted evaluation methods have paved great promises for the systematic, timely and through assessment of learning outcomes. Though many techniques from the computing paradigm are put to practice, 'Educational Data Mining' is the one which has been less explored in spite of its advantage in terms of immediate feedback offered to the teachers. The paper portrays an evaluation model based on the cluster data analysis to sequence and re-sequence the order of the multiple choice questions with the increasing level of difficulty. The paper reveals a novel evaluation paradigm based on the record keeping algorithm implemented on the computer to look at the sequence of questions attempted by the learner and time spent on each question. Online multiple choice test serves as the input for the system while the sequence of the attempted questions vis-à-vis time spent on the respective question serves as the system metrics. Similarity based clustering that forms as an inherent implementation technique has been dealt in depth in the paper.

Findings :

Cluster having more number of data sets gives sequence followed by most of the students while attempting the test. The time spent on each question determines the complexity of questions. This also assists in strategizing the teaching strategy which is focused on giving more emphasis on the weaker areas of the students. The methodology is also useful for the self analysis of the faculty member's inorder to perceive where their teaching is really not up to the mark.

Conclusions :

The paper showcases application of similarity based clustering for analyzing the progression of the students by using the multiple choice examinations as a tool. Cluster having more number of data sets gives sequence followed by most of the students while attempting the test. Using this data question paper can be *re-sequenced so that the students would face the questions with gradually increasing level of difficulty which in turn will boost their confidence. Thus the computer assisted methodology based on similarity clustering method opens yet another dimension to make the MCQ type examinations more effective.*

Keywords: *cosine similarity, clustering, question sequencing, educational data mining*

INTRODUCTION

Multiple-choice questions (MCQs) are being increasingly used in higher education as a means of supplementing or even replacing current assessment practices. The growth in this method of assessment has been driven by wider changes in the higher education environment such as the growing numbers of students, reduced resources, modularization and the increased availability of computer networks (Nicol, 2007). The other natural advantage of the MCQs in assessing the learning outcome is the inherent suitability for automating the evaluation process thereby removing the teacher's personal bias towards the learners. Therefore even in the developing countries like India, the MCQs are gaining wide popularity and in good number of instances such as the entrance examinations, competitive examinations for various government jobs, they have poised to become the sole evaluation strategy. However the ease of implementation with the widespread digital assisted evaluation, there are several concerns favoring the MCQs as the sole evaluation strategy. Though the faculty members in the institutes of higher learning are resorting to standard techniques of paper setting such as Bloom's taxonomy, there has been less attention towards the subsidiary data generated while the learner solves the question paper. The massive amount of subsidiary data which has the potential to throw light on the hidden dimensions of the learning outcomes which might be useful to formulate the pedagogical strategies is simply remains unutilized in the while process. Data Mining has a big role to play here, to extract the hidden, unknown and potentially useful information and patterns from the above mentioned data repositories. This has led to 'Educational Data Mining' as the discipline in itself to provide an insight into student's learning patterns and the environment in which they learn. It can also help in designing the curricula, scheduling the classes, providing recommendations for students as well teachers and so on.

International scenario in this context reveals many researchers striving hard to use different tools and techniques from the computing paradigm for enhancing the effectiveness of the evaluation techniques. A fuzzy based tool is developed (Hui, Duo, Mingli, & Lei, 2010; Malvezzi, Mourão, & Bressan, 2010; Weon & Kim, 2001) to follow up and to evaluate the student learning which enable the coordination of Classroom mediated by technology (CMT) courses to elect the best teaching practices that allow the increasing of student's performance. A classifier is designed (Minaei-Bidgoli, Kashy, Kortmeyer, & Punch, 2003) to classify Students based on features extracted from logged data in an education web-based system to predict their final grades. It was observed in (West, 2012) that there is potential for improved research, evaluation, and accountability through data mining, data analytics, and web dashboards. The "big data" make it possible to mine learning information for insights regarding student performance and learning approaches. Rather than rely on periodic test performance, instructors can analyze what students know and what techniques are most effective for each pupil. Application of data mining in education for student profiling using Apriori algorithm and student grouping using K-means clustering is carried out by (Parack, Zahid, & Merchant, 2012). A measure to compute similarity between sequences containing accesses to Web pages, and a centroid-based clustering approach for grouping sessions of accesses to a Web site has been proposed by (Manco, Ortale, & Saccà, 2003). A new method for students' learning achievement evaluation by automatically generating the importance degrees of the attributes of questions has been proposed by (S.-M. Chen & Li, 2010). Similarly the AHP (Analytic Hierarchy Process) method's basic theory analysis is seen to be used in online-teaching platform, analyzing ability of students' self-learning, and establishing the evaluation index system by (Y. Chen & Yang, 2010). In an another effort a framework for analyzing clustering from similarity information that directly addresses this question of what properties of a similarity measure are sufficient to cluster accurately and by what kinds of algorithms according to expert criteria derived the value of influence factors has been discussed in (Balcan, Blum, & Vempala, 2009).

In the backdrop of the above trends, the present paper reports application of the similarity based clustering for augmenting the learning outcome as well as stargazing the teaching methodology. The paper is divided into several sections. After introducing the theme, the basics of similarity based clustering technique is presented. This is followed by applying the technique for a sample of ten students. The results have been analyzed in view of learning outcomes and identifying the gaps in the teaching learning process.

APPLYING THE SIMILARITY BASED CLUSTERING APPROACH FOR MCQ EVALUATION

Similarity based clustering methods are very effective and robust approaches for clustering ‘big data’ on the basis of a total similarity objective function related to the approximate density shape estimation. These methods find widespread applications in diverse application domains, including biomedical problems, but also in remote sensing, geoscience or other technical domains (*Similarity-Based Clustering - Recent Developments and Biomedical Applications*, n.d.).

In the present work, in order to apply the similarity based clustering; the profile of the students is gathered prior to the examination. The above said profile describes the student’s interests, preferences and time requirements for solving a question. In order to identify the sequence of MCQs attempted by the student following metrics were extracted:

- Sequence of questions attempted by the student
- Time spent on each question
- First question solved by the student
- Number of attempts to a particular question
- Overall time taken by student to solve the entire paper

The analytical treatment to come out with the Similarity based Clustering based on the above metrics goes on the following lines:

Let $q_s=(q_{i1},q_{i2},\ldots\ldots\ldots q_{in})$ is the vector set that defines sequence of questions attempted by the student and let $t_s=(t_1,t_2,\ldots\ldots\ldots.t_m)$ is the vector set that defines sequence of time spent on particular question.
Then set $S=\{(q_{i1},t_1),(q_{i2},t_2),\ldots\ldots..(q_{in},t_m)\}$ is the vector set that defines question attempted by the student and time spent on that particular question.

Considering three students attempting exam of ten objective questions, question sequence of three students is (q_{s1}, qs2, qs3)

$q_{s1}=(q_1,q_2,q_3,q_4,q_5,q_6,q_7,q_8,q_9,q_{10})$
$q_{s2}=(q_1,q_2,q_3,q_4,q_5,q_6,q_7,q_8,q_9,q_{10})$
$q_{s3}=(q_1,q_2,q_3,q_4,q_5,q_6,q_7,q_8,q_9,q_{10})$

Each above sequence has corresponding time sequence that is time spent by student on each question.

$q_{ts1}= (t_1$, t2, t3)
$q_{ts1}= (t_1$, t2, t3)
$q_{ts1}= (t_1$, t2, t3)

Similarity between students can be computed as follows ,

$$\mathrm{sim}^s(s_i,s_j)=\begin{cases} 1 & \text{if } i=j \\ q_s\mathrm{sim}_{qus}(q_{si},q_{sj})+t_s\mathrm{sim}_{time}(t_{si},t_{sj}) & \text{if } i\neq j \end{cases} \quad \ldots\ldots\ldots\ldots\ldots (1)$$

where, sim_{qus} refers to question similarity and sim_{time} refers to time spent on each question similarity.
And $q_s+t_s=1$ to quantify both question and time sequence
The term $\mathrm{sim}_{qus}(q_{si},q_{sj})$ computes the cosine similarity between the vectors associated to the questions attempted by the student $s_i$and s_j.

$$\mathrm{sim}_{qus}(\overrightarrow{q_{si}},\overrightarrow{q_{sj}}) = \frac{\overrightarrow{q_{si}}\,\overrightarrow{q_{sj}}}{\|\overrightarrow{q_{si}}\|\|\overrightarrow{q_{sj}}\|}$$

$$= \frac{\sum_{i,j=1}^{n} q_{si} \times q_{sj}}{\sqrt{\sum_{i=1}^{n} q_{si}^{2}} \times \sqrt{\sum_{j=1}^{n} q_{sj}^{2}}} \quad \ldots\ldots\ldots\ldots\ldots (2)$$

The resulting similarity ranges from -1 (exactly opposite) to 1(exactly same) and 0 indicates independency and in-between values indicates intermediate similarity or dissimilarity.

$sim_{time}(t_{si},t_{sj})$ can be computed using cosine similarity formula (2) as above.
The formula (1) gives the similarity factor between students considering both question sequence and time spent on each question.

Algorithm was developed and the software suite was developed in a LAN environment. The methodology for analysis and other details are presented in the following section.

VALIDATION AND TEST BED GENERATION

Similarity based Clustering as applied to the MCQ evaluation was validated for a data sample of a class of 40 students. The analysis methodology for a small subset of the ten students is presented here.

Data set of ten students each attempting ten MCQs is as shown in table 1. The data is further clustered in to 3 clusters namely C_1, $C_2$and C3.

Table 1: Test Data Set

Student	Sequence of Questions attempted									
S_1	1	2	3	5	6	8	9	7	4	10
S_2	1	3	2	5	6	7	9	8	4	10
S_3	2	4	10	9	8	7	6	5	1	3
S_4	10	1	2	3	4	5	6	7	8	9
S_5	1	2	3	4	5	6	7	8	9	10
S_6	1	2	9	10	7	6	8	5	4	3
S_7	1	2	3	4	5	6	7	8	9	10
S_8	1	5	4	9	3	6	7	8	10	2
S_9	2	3	4	5	6	1	7	8	9	10
S_{10}	1	2	3	4	7	8	9	10	5	6

From the above three clusters, randomly assigned centroids were calculated as given below:

S_1, S_5, $S_8$are the names associated with the randomly assigned centroids for C_1, C_2 and C_3.

C_1:

S_1	1	2	3	5	6	8	9	7	4	10

C_2:

S_5	1	2	3	4	5	6	7	8	9	10

C_3:

S_8	1	5	4	9	3	6	7	8	10	2

Now using formula (2) similarity measure of each data set with respect to $C_1$was calculated and the same is presented in table 2.. Following table gives similarity measure related to C_1.

Table 2: Similarity measure for cluster C_1

Student	Computed Value
S_2	0.99
S_3	0.81
S_4	0.83
S_6	0.84
S_7	0.95
S_9	0.89
S_{10}	0.96

Following the same procedure, similarity measure for C_2 with respect to all the data set was calculated and the same is presented in table 3.

Table 3: Similarity measure for cluster C_2

Student	Computed Value
S_2	0.95
S_3	0.72
S_4	0.88
S_6	0.79
S_7	1.00
S_9	0.96
S_{10}	0.93

Considering C_3 and Formula (2) to calculate similarity measure for cluster C3 was calculated as given in table 4.

Table 4: Similarity measure for cluster C_3

Student	Computed Value
S_2	0.82
S_3	0.79
S_4	0.74
S_6	0.87
S_7	0.86
S_9	0.84
S_{10}	0.86

Comparing the computed values from Table 2, Table 3 and Table 4 , the clusters describing similarity between students were calculated and the same is presented in table 5.

Table 5: Resulting Clusters

C_1	C_2	C_3
S_1	S_5	
S_2	S_4	S_8
S_3	S_7	S_6
S_{10}	S_9	

Thus for the cluster C1 of the four students, the re-sequencing of the questions is required as S1, S2, S3 and S10. The same suite is followed for clusters C2 and C3. The methodology was actually tested for a multiple choice question paper for MCA students on the subject "Theory of Computation". Details are presented in the following section.

CLUSTER DATA ANALYSIS AS APPLIED TO "THEORY OF COMPUTATION"

Multiple choice based test was conducted for Master in Computer Applications (MCA) students of the Shivaji University, Kolhapur on the subject "Theory of Computation". It was online test with ten questions. Time allotted for the test was 10 minutes i.e one minute per question. Questions were based on following topics in a sequence:

1. Finite automata
2. Finite automata with output
3. Regular expression
4. Minimization of DFA
5. Arden's theorem
6. CFG
7. GNF
8. PDA
9. NPDA
10. Turing Machine

Depending on the sequence of question attempted by the students and the amount of time spent on a particular question, students were categorized in to three clusters as shown in the table 5. Following were the observations.

Students in cluster- one (C_1) spent an average of 0.30 seconds per question, and almost all of them attempted first question first i.e. all of them understood the topic finite automata well and are confident about it. . Number of attempts by the students for question number 9 were more as compared to other questions. This indicates students in this cluster have difficulty in understanding Non-deterministic Pushdown Automata(NPDA), so extra coaching on this topic should be provided to improve their results.

Average time spent on each question by students in cluster-2 (C_2) was 0.50 seconds. Approximately all the students i.e three out of four students attempted first question first again showing their understanding in the topic finite automata. Question no. 3 and 9 had more number of attempts indicating students understanding level in Regular expression and Non- deterministic Pushdown Automata(NPDA).

Cluster-3 (C_3) had all the students who were weak in most of the topics. Time spent by them on some questions like question no 2, 5,6, 7, 9 was more as compared to other questions and also all these questions had more number of attempts. This indicates students in this cluster have not understood 50% of the syllabus and needs rigorous coaching.

Over all observation is topic Non-deterministic PDA needs special attention. We need to analyze whether question number 9 needs to reframe or the topic needs to be elaborated more. Students in cluster-1 are doing well in their academics and can be graded as good students and students in cluster-2 are average students though time spent in average per question is 0.50 seconds but result is not good. Cluster -3 students' needs to put in lots of hard work to improve their academics and also instructors need to pay special attention to them.

As per the sequence of question paper cluster-1 and cluster-2 students are good enough with it with some difficulty in question no-3 pertaining to cluster-2 and question no. 2, 5,6,7,9 pertaining to cluster-3. These questions need to be reframed depending on the result.

CONCLUSION

The paper showcases application of similarity based clustering for analyzing the progression of the students by using the multiple choice examinations as a tool. Cluster having more number of data sets gives sequence followed by most of the students while attempting the test. Using this data question paper can be re-sequenced so that the students would face the questions with gradually increasing level of difficulty which in turn will boost their confidence. The time spent on each question determines the complexity of questions. This also assists in strategizing the teaching strategy which is focused on giving more emphasis on the weaker areas of the students. The methodology is also useful for the self analysis of the faculty member's in order to perceive where their teaching is really not up to the mark. Thus the computer assisted methodology based on similarity clustering method opens yet another dimension to make the MCQ type examinations more effective.

REFERENCES

Balcan, M.-F., Blum, A., & Vempala, S. (2009). Clustering via Similarity Functions: Theoretical Foundations and Algorithms. *Volume: V, Publisher: Citeseer, Available from citeseerx. ist. psu. edu*. Retrieved from http://www-cgi.cs.cmu.edu/afs/cs.cmu.edu/Web/People/avrim/Papers/BBVclustering_journal.pdf

Chen, S.-M., & Li, T.-K. (2010). A new method to evaluate students' learning achievement by automatically generating the importance degrees of attributes of questions. In *Machine Learning and Cybernetics (ICMLC), 2010 International Conference on* (Vol. 5, pp. 2495–2499). Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5580818

Chen, Y., & Yang, M. (2010). Study and construct online self-learning evaluation system model based on AHP method. In *Information and Financial Engineering (ICIFE), 2010 2nd IEEE International Conference on* (pp. 54–58). Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5609317

Hui, Y., Duo, L., Mingli, Y., & Lei, S. (2010). Evaluation and Study of Students Automatic Learning Under Network Environment. In *Information Engineering (ICIE),* *2010 WASE International Conference on* (Vol. 2, pp. 84–86). Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5571254

Malvezzi, W. R., Mourão, A. B., & Bressan, G. (2010). Learning evaluation in Classroom mediated by technology model using fuzzy logic at the University of Amazonas State. In *Frontiers in Education Conference (FIE), 2010 IEEE* (p. S2C–1). Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5673494

Manco, G., Ortale, R., & Saccà, D. (2003). Similarity-based clustering of Web transactions. In *Proceedings of the 2003 ACM symposium on Applied computing* (pp. 1212–1216). Retrieved from http://dl.acm.org/citation.cfm?id=952767

Minaei-Bidgoli, B., Kashy, D. A., Kortmeyer, G., & Punch, W. F. (2003). Predicting student performance: an application of data mining methods with an educational web-based system. In *Frontiers in Education, 2003. FIE 2003 33rd Annual* (Vol. 1, p. T2A–13). Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1263284

Nicol, D. (2007). E-assessment by design: using multiple-choice tests to good effect. *Journal of Further and Higher Education*, *31*(1), 53–64.

Parack, S., Zahid, Z., & Merchant, F. (2012). Application of data mining in educational databases for predicting academic trends and patterns. In *Technology Enhanced Education (ICTEE), 2012 IEEE International Conference on* (pp. 1–4). Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6208617

Similarity-Based Clustering - Recent Developments and Biomedical Applications. (n.d.). Retrieved from http://www.springer.com/computer/bioinformatics/book/978-3-642-01804-6

Weon, S., & Kim, J. (2001). Learning achievement evaluation strategy using fuzzy membership function. In *Frontiers in Education Conference, 2001. 31st Annual* (Vol. 1, p. T3A–19). Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=963904

West, D. M. (2012). Big Data for Education: Data Mining, Data Analytics, and Web Dashboards. *Governance Studies. Brookings, US: Reuters*. Retrieved from http://www.brookings.edu/~/media/Research/Files/Papers/2012/9/04%20education%20technology%20west/04%20education%20technology%20west.pdf